NAS Technical Report: NAS-2014-02

Supporting "Big Data" Analysis and Analytics at the NASA Advanced Supercomputing (NAS) Facility

Piyush Mehrotra
L. Harper Pryor[1]
F. Ron Bailey[1]
Marc Cotnoir[1]

NASA Advanced Supercomputing (NAS) Division,
NASA Ames Research Center, Moffett Field, CA 94035

piyush.mehrotr@nasagov, l.harper.pryor@nasa.gov,
frank.r.bailey@nasa.gov, marc.cotnoir@nasa.gov

Abstract

The NASA Advanced Supercomputing (NAS) Division is the leading provider of computation and related services for the NASA engineering and scientific simulation community. As part of a continuous process to understand and anticipate the changing needs of the NAS user community and expand support in computational sciences to support key NASA goals and initiatives, we interviewed 12 individuals representing both the NAS user community and other researchers who could provide insight into emerging needs. In particular, we focus on challenges related to the rapidly growing area of "big data" analytics, which affects both in the scientific and business communities. We also seek to understand other areas where NAS's core competency in computational sciences can support key NASA goals and initiatives. This report presents the background, questions, processes, key findings from an examination of the impact of these issues, and identifies next steps NAS's next steps, both now and in the future.

January 29, 2014

Version 1

[1] Computer Sciences Corp., NASA Contract NNA07CA29C

`

1. Statement of the Problem & Background

The NASA Advanced Supercomputing (NAS) Division is well established as the leading provider of computational resources and related services for NASA's engineering/scientific modeling and simulation communities. To maintain this position, NAS constantly strives to understand and anticipate the needs of these communities in order to evolve NAS services to meet their changing needs. Division leaders also seek to understand other areas where the NAS core competency in computational sciences can support key NASA goals and initiatives.

An important issue facing NASA, for both current NAS users and potential new users, is the explosion of data that is variously referred to as "big data" and "data-intensive science." NAS users are already involved in the latter. Their codes use and generate very large datasets, and analysis of these datasets is an important part of the scientific/engineering workflow. NASA is also an important provider of big data, particularly satellite remote sensing data. The challenge of extracting knowledge and information from such large datasets is driving the emergence of new approaches to "big data analytics" and "predictive analytics," both in the scientific and business communities.

This report presents the results of an examination of the impact of these issues on NAS. Specifically, we set out to address the following questions:

- What does NAS need to do to serve the analysis/analytics needs of our users—now and in the future?
    - Does Pleiades fill the need or does NAS need to do something else?
- How we can expand NAS's role in analysis/analytics of NASA big data?
    - Can we provide similar services to the big data community as we provide to the simulation community?

To answer these questions, we need to:

- Understand what our users are doing and where they are going in analysis/analytics
- Understand what others are doing and where they are going in analysis/analytics
- Determine the unmet needs

2. Definitions

Big data, *analysis*, and *analytics* are not precisely defined terms, but some definition is needed to provide context.

Big data is defined by more than just the amount of data. The key point about big data is that the size and characteristics of datasets being addressed overwhelm traditional, existing data management and analysis techniques, and therefore require novel algorithms, infrastructure, and frameworks to perform advanced analytics. Big data is data processing at a scale that is not just quantitatively big, but qualitatively different. This is similar to some informal definitions of supercomputing which state that if you can do it with mainstream technology, it isn't supercomputing.

A common way of characterizing big data is according to "three Vs" namely volume, velocity, and variety. The point being that, in addition to sheer size, the speed of data moving across systems and networks and the variety of content add to the challenge of exploiting big data.

Analysis and analytics are best thought of as a continuum, as shown below. At one end of the spectrum, analysis is characterized by more knowledge about the data and the processing to be applied, while at the other end of the spectrum, analytics is characterized by less knowledge about the data and the end result being sought. Another way to summarize the difference is that *analysis is more about interpretation*, whereas *analytics is more about exploration.*

Analysis vs. Analytics A Continuum from the Known to the Unknown		
Analysis		**Analytics**
You know what you want to know, where to look, and how to find the answer	⇔	***You don't know what you what to know, where to look, or how to search for it***
Almost always looking for quantitative results	⇔	**Often looking for qualitative results**
Searching for parameters	⇔	**Searching for relationships**
More numeric than semantic		**More semantic than numeric**
Applied to specific datasets to find specific information	⇔	**Applied to non-specific datasets to find unknown information**
INTERPRETATION	⇔	**EXPLORATION**

3. Approach

The primary information source for this examination of analysis/analytics needs is a set of interviews with a cross section of individuals representing both the NAS user community and others working in analysis/analytics who could provide insight into emerging needs. We also conducted a brief search for big data initiatives and activities within NASA. In identifying individuals to interview, the goal was to span multiple application areas and to cover a range of different user organizations. We also sought to include researchers "pushing the envelope" on analytics. The full list of individuals interviewed is provided in Appendix I.

Before conducting the interviews, we developed a Data Analytics Framework to provide context for the information we would gather. (See Appendix II, which also discusses how NAS fits into this framework.) We then developed a questionnaire that was used to guide each interview. In general terms, the interview sought to determine what data the individual was using (including where the data reside and what data transfers are needed), what analysis/analytics were being performed (including algorithms and tools and where the processing is done), and what the interviewee envisioned as future needs. We asked for quantitative information about the size of the analysis problems (such as the amount of data, amount of computation) but interviewees had very little quantitative information. The full questionnaire is presented in Appendix III.

Most of the interviews were conducted with two NAS team members participating, although for schedule reasons a few were conducted with only one team member. After each interview, each team member captured notes independently, and then these notes were combined. Having two team members on the calls and combining notes this way was valuable to make sure we had full understanding of interviewee responses.

After the interviews were complete, each team member reviewed all of the notes and documented findings and observations. These were combined into a presentation that was delivered on July 18, 2013. Following discussion of the presentation, the definition of the implications for NAS system planning were further refined.

4. Current NASA Big Data Efforts

In addition to conducting interviews, we searched for information about current big data work within NASA. This search revealed the following:

- While there is interest in big data throughout NASA, there is no single NASA big data initiative.
- There is some confusion between the NASA Open Data project (part of the NASA Open Government Initiative) and big data. The Open Data project is about visibility and access. Their high-level index points to existing data access capabilities—for example in the Earth Observing System Data and Information System (EOSDIS)—but there do not appear to be any new capabilities coming out of this initiative at this time, and in particular nothing that addresses big data. The NASA Earth Exchange (NEX) and DASHlink are mentioned as examples of activities targeted at making data, algorithms, and research results more easily available to the research community.
- Discussions and presentations on big data within NASA all tend to identify and feature the same handful of projects, such as the efforts to process lunar mapping data and the aircraft safety project at Ames. They identify various Announcements of Opportunity that include or have included big data type topics. Within NASA's Science Mission Directorate (SMD), mention is made of access to NASA satellite data via the Distributed Active Archive Centers (DAACs) and NEX. NEX is often featured as a major thrust aimed at expanding use of NASA datasets.
- While NASA was not one of the six lead agencies included in the March 2012 Obama Administration "Big Data Research and Development Initiative," the supporting press release cited NASA activities including the Advanced Information Systems Technology (AIST) Program, Earth Science Data and Information System (ESDIS) Project, Global Earth Observation System of Systems (GEOSS) effort, the Planetary Data System (PDS),

the Mikulski Archive for Space Telescopes (MAST), and the NASA Earth Science Gateway (ESG).

- The community that has done the most with big data at NASA is the Earth science community. This includes work at the DAACs for ESDIS, the NASA Center for Climate Simulation (NCCS), and NEX. The best routes into this work are the Earth Science Data System Working Groups (ESDSWGs) and the Federation of Earth Science Information Partners (ESIP), which are closely related.
- The NASA centers most involved with big data are Goddard Space Flight Center, Marshall Space Flight Center, Langley Research Center, and the Jet Propulsion Laboratory.

Note that beyond NASA, there is relevant work within other agencies that NAS will want to review for applicable ideas, in particular the National Science Foundation, National Institute for Science and Technology, Department of Energy, U.S. Geological Survey, National Oceanic and Atmospheric Administration, Department of Defense, and National Institute of Health.

5. Key Findings from the Interviews

Summary of Findings

The following are the major findings from analysis of the interviews and subsequent discussions. Each is discussed briefly below:

- Today, NAS users do a lot more analysis than analytics.
- Users perform a broad range of algorithms and processes on data.
- Many data analysis tools are user developed.
- Nearly all applications involve large observational datasets.
- Data is structured in most cases.
- Nearly all datasets are many terabytes (TBs) in size, with some reaching a few petabytes (PBs).
- Most of the analysis/analytics processing is done at NAS or the NCCS.
- The large-scale datasets used for analysis/analytics reside at NAS or the NCCS.
- Users want easy access to data.
- Big data analysis/analytics requires large-memory configurations and high-bandwidth I/O but is not computationally intensive.

Discussion

- **Today, NAS users do a lot more analysis than analytics**. While this is true of most of the interviewees, two interviewees (Kumar and Oza) were performing work that is clearly at the analytics end of the continuum. Nearly all interviewees said they are doing at least some exploratory analysis. Note that this result is influenced by the particular set of interviewees and it is possible that we have not found the people doing new and different kinds of analytics on NASA datasets.

- **Users perform a broad range of algorithms and processes on data**. The table below lists those that were mentioned. Consistent with the above, most of these would be considered analysis, where a known algorithm/process is applied to well-understood data. The items toward the bottom of the list are more analytic (searching for relationships and machine learning).

Algorithms and Processes mentioned by Interviewees

- **Statistical analysis**
- **Time series analysis**
- **Eigen decomposition**
- **Iso-surface extraction**
- **Feature detection/extraction**
- **Structure identification**
- **Line tracing**
- **Multivariate analysis**
- **Subsetting and filtering**
- **Change detection/characterization**
- **Signal processing**
- **Search for relationships**
- **Machine learning algorithms**

- **Many data analysis tools are user developed**. These may be fully custom or semi-custom software built with packages/libraries and scripting. Interviewees also reported the use of packaged tools, including both general mathematical packages and libraries and application-specific packages. MATLAB is heavily used. The following table shows tools that were mentioned in the interviews.

Tools mentioned by Interviewees

- **MATLAB**
- **IDL (GDL)**
- **ENVI**
- **Tecplot**
- **Python**
- **ParaView**
- **LEDAPS**
- **FieldView**
- **Geosail Flight**
- **GrADS**
- **GEMPAK**
- **METS**
- **NCL**

- **Nearly all applications involve large observational datasets**. Those applications that involve simulation models often incorporate observed data either for data assimilation or for comparison to model results. The following table shows specific types of datasets mentioned by interviewees.

**Types of Datasets
mentioned by Interviewees**

- **Earth Science data**
  - **Satellite data**
  - **Model output**
  - **Other observational data**
  - **Ancillary datasets (e.g., DEM)**
- **Aeronautics data**
  - **Simulation output**
- **Other domain specific data**
  - **Kepler telescope data**
  - **Flight recorder data**

- **Data is structured in most cases**. There were a few cases of less structured data, such as datasets from flight data recorders or point observational data, but in this sample, these were the exceptions.

- **Nearly all datasets are many TBs in size with some reaching a few PBs**.

- **Most of the analysis/analytics processing is done at NAS or the NCCS**. At NAS, interviewees mentioned Pleiades, Endeavour, and NEX. Some processing is done on users' local systems (compute clusters or workstations), but interviewees stated that limitations in storage and networking limit what can be done locally. This is especially true for visualization. Basically, they do the analysis where the data are, which leads to the next point.

- **The large-scale datasets used for analysis/analytics reside at NAS or the NCCS**. When the source is elsewhere, such as DAACS or the Program for Climate Model Diagnosis and Intercomparision (PCMDI), the data must be moved to the location where the processing will be done.

- **Users want easy access to data**. They do not want to have to move data around. If they have to, they want it to be easy.

- **Big data analysis/analytics requires large-memory configurations and high-bandwidth I/O but is not computationally intensive**. Unlike the large-scale engineering and simulation codes, analysis and analytics is not computationally intensive. In fact, many of these applications today run on single processors, although parallel applications are emerging and will probably grow. When parallel techniques are used, interviewees cite reasons of parallel access to data, not for access to more compute power. (This is the case for the MapReduce paradigm, for example.)

6. Use Cases and Common Processes

As part of developing the implications of these findings for NAS, we identify use cases for big data analysis and analytics on NASA data. These use cases will inform our thinking about possibilities for NAS to respond to the challenges users face in executing the use cases. This set of use cases attempts to represent the types of analysis and analytics being performed by the interviewees, spanning the continuum from analysis to analytics and representing various types of users.

The set of use cases discussed here assumes all the processing is performed at NAS on data that is resident at NAS. However, in other real-world cases, it may be necessary to move the datasets to NAS for processing. The datasets can often be NASA data but may be datasets from other sources. Likewise, it may be preferred to move results after processing or to provide a means to disseminate the results.

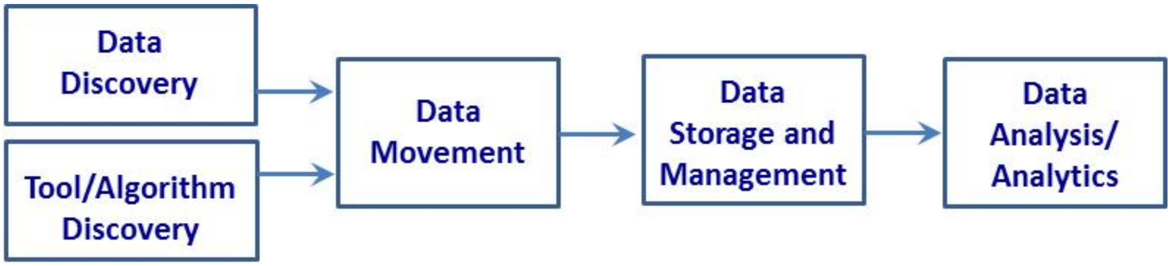
From the use cases, we can identify a set of common processes that users must be able to execute to accomplish the use cases, as well as the challenges associated with executing these processes.

The use cases considered are listed here. They are discussed in more detail in Appendix IV.

- **User Goal: Produce a derived dataset by processing NASA data**.
- **User Goal: Find NASA data relevant to a scientific problem**.
- **User Goal: Discover new characteristics/features in a NASA dataset**.
- **User Goal: Assess the goodness of a simulation dataset**.
- **User Goal: Answer a scientific question through analysis of or analytics on NASA data**.
- **User Goal: Provide the results of analysis/analytics to others**.

From the use cases, a set of common processes that users need to be able to perform is abstracted, and the challenges faced by users in executing the use cases can be associated with these common processes to form the basis for implications for the NAS roadmap for supporting big data analysis and analytics.

Five common processes appear across the use cases. The five processes define a top-level workflow for analysis/analytics as shown in the following figure.



The following are the identified challenges users face in executing these processes.

- **Data Discovery: A user wants to discover datasets applicable to a scientific problem**.
  - The user needs a way to discover what datasets exist and where the datasets are located. This is made difficult by the fact that data are distributed across many sites and there is great diversity in the types of data available.
  - The user needs a way to discover the characteristics of the data in the datasets as they relate to the scientific problem.
  - The user needs a way to discover the characteristics of the data in the datasets as they relate to accessing and manipulating the data.
- **Tool/Algorithm Discovery: A user wants to discover tools/algorithms applicable to a scientific problem**.
  - The user needs a way to discover what tools exist, where they are, and how to access them. This is made difficult because there is no standard nomenclature or metadata for tools/algorithms.
  - The user needs a way to assess the applicability of tools/algorithms to the specific problem to be solved.
- **Data Movement: A user wants to move potentially very large datasets from another site to NAS or from NAS to another site**.
  - Requires a user-friendly transfer mechanism—tools to make it easy to accomplish data movement.
  - Requires adequate network bandwidth.
  - The user is often faced with having to understand details of the environment at both the source and destination site.
  - The transfer could be one time or could be ongoing.
- **Data Storage and Management: A user wants to store large amounts of data and manage access to it**.
  - Requires large filesystems with high I/O performance.
  - Requires the ability to make metadata visible to users.
  - Requires the ability to access data in many different formats.
- **Data Analysis/Analytics: A user wants to execute an algorithm against a large (TB scale) dataset**.
  - Requires a platform with large memory space and high I/O bandwidth.
  - Necessary datasets have to be available “close” to the computational platform.
  - Requires large storage space for datasets, both the source datasets and results.

7. Implications for NAS and Recommendations

As stated in the introductory section, the purpose of this study has been to address two questions:

- What does NAS need to do to serve the analysis/analytics needs of our users – now and in the future?
  - Does Pleiades fill the need or does NAS need to do something else?
- How we can expand NAS' role in analysis/analytics of NASA big data?
  - Can we provide similar services to the big data community as we provide to the simulation community?

The following discussion provides the implications for some of what was learned. Section 7.1, "Architecture/Environment Roadmap" addresses the first question regarding steps NAS needs to take to serve the analysis/analytics needs of users, along with a brief discussion of NAS' current capabilities and resources in each area. The potential role of Pleiades is addressed in these discussions.

Sections 7.2 and 7.3, "User Community" and "Role of NEX," respectively, address possibilities for expanding NAS's role in analysis/analytics of NASA big data. Finally, Section 7.4, "Path Forward" recommends specific steps that can be taken to advance an analysis/analytics initiative at NAS.

7.1 Architecture/Environment Roadmap

The key factor in supporting big data analysis and analytics is that users need to bring together the tools and the data in an environment that supports analysis/analytics processing. Meeting this need effectively has several implications for the NAS environment. Specifically, to support big data analysis and analytics on NASA's big data, NAS must:

- Provide users with easy access to a variety of potentially very large datasets at NAS. This will mean petabyte-scale storage.
- Make it easy for users to move data to and from NAS. This means, at a minimum, from the DAACS and to/from users at other NASA centers.
- Provide users with a rich set of tools/algorithms that span the range from analysis to analytics, as well as the environment to develop their own tools.
- Provide users with computational platforms that support large memory spaces and have very high I/O bandwidth.

Each of these is discussed briefly below.

- **Provide users with easy access to a variety of potentially very large datasets at NAS.** This will mean petabyte-scale storage. As part of NAS support to NEX, NAS provides the infrastructure for storing and accessing large Earth science datasets and hosts a collection of NEX "core datasets." NAS filesystems also accommodate very large output datasets from simulations in aeronautics and in Earth and space science. Storage assets include both disk storage and archive (tape) storage, along with tools to manage migration across tiers of storage. To provide ease of access from across platforms, NAS has been developing the infrastructure to allow sharing of the datasets. In the future, both the total amount of data that will need to reside at NAS and the size of individual datasets is expected to grow, and these capabilities will need to scale to accommodate this growth.

- **Make it easy for users to move data to and from NAS**. This means, at a minimum, from the DAACs and to/from users at other NASA centers. NAS networking staff work with users at remote sites to solve networking problems and increase the bandwidth realized for transfers to/from NAS. In addition, NAS-developed tools, including the Secure Unattended Proxy and SHIFT, have been implemented and documentation is available in the HECC Knowledge Base. Users have expressed a need for more support and tools for data movement.

- **Provide users with a rich set of tools/algorithms that span the range from analysis to analytics, as well as the environment to develop their own tools**. NAS endeavors to provide the both the software tools that users request on NAS computational platforms, including licenses to heavily used tools, and an environment for users to develop their own tools. Based on user feedback, NAS expects the number and diversity of tools and algorithms to increase, as indicated by the range of tools and algorithms mentioned above in Section 5. NAS needs to examine additional ways to provide visibility so that users can discover capabilities that are available on NAS platforms. This is one of the specific goals of the NEX collaborative environment hosted at the NAS facility.

- **Provide users with computational platforms that support large memory spaces and have very high I/O bandwidth**. NAS offers a variety of platforms that are suitable for analysis/analytics, including the Pleiades bridge nodes, the Lou data analysis nodes and the Endeavour system. These platforms provide either larger-memory nodes or shared-memory environments, which are the capabilities stated by users as necessary for analysis/analytics and that are used for post-processing of simulation results. The nature of this post-processing is generally at the analysis end of the analysis-to-analytics continuum described in Section 2. NAS has additional platforms such as the Merope cluster and a many-integrated-core (MIC)-based test system, Maia, whose applicability is uncertain at this point. Further quantification of the computational and I/O demands for analysis/analytics, based on elaboration of use cases, is needed to determine a specific platform roadmap.

- **Pleiades is optimized for large-scale simulation, however as noted, the bridge nodes are configured for post-processing**. The flexible architecture of Pleiades allows for subsets of nodes to be optimized for different workloads, so evolution of the bridge nodes or the addition of other nodes with configurations tailored for analysis/analytics is one architectural path that NAS can follow. The close coupling of such nodes with the rest of Pleiades and with storage assets via high-speed networking offers possibilities for integrating analysis/analytics with simulation, which is one of the trends related to large-scale analysis/analytics.

7.2 User Community

The interviews conducted for this study focused mostly on the current NAS user community and related applications. By providing effective support for analysis and analytics, NAS will expand the services provided to existing users—however NAS could also attract new users. NAS should explore outreach opportunities to begin to identify these users and make them aware of the division's capabilities and services.

7.3 Role of NEX

In these interviews, it became clear that there is a strong connection between the needs of users for analysis/analytics and the goals of NEX to "enable enhanced and more efficient use of Earth observations for NASA Earth science technology, research, and applications programs." NEX is specifically exploring capabilities for discovery of data and modeling resources— one of the need areas identified for analysis/analytics—and NEX is obtaining a range of Earth science datasets, including satellite data and climate datasets. While Earth science is only one application domain served by NAS, it is clearly the domain with the largest big-data assets and challenges.

7.4 Path Forward

The following are specific next steps to advance a NAS position in big data analysis and analytics.

- **Continue Interviews**. It would be valuable to conduct additional interviews with an expanded set of individuals from outside the traditional NAS user community, in order to find the leading-edge researchers and practitioners in big data for scientific and engineering applications. These individuals could be involved with NASA applications or non-NASA applications that could gain value from exploring NASA big data.
- **Refine and quantify the use cases**. This will provide quantification of the demands on computational and storage resources. The first use case to elaborate on would be, "Answer a scientific question through analysis of or analytics on NASA data."
- **Define a NAS Analysis/Analytics Environment**. This will clarify NAS capabilities and support both planning to evolve the environment and outreach to current and potential users. The environment can be described in terms of NAS services for users performing analysis/analytics. Note that NEX is one model for presenting a set of capabilities to users.
- **Collaboration**. NAS should establish a dialog with other NASA organizations involved with big data initiatives, including the NCCS, the NASA Land Processes Distributed Active Archive Center, and the Atmospheric Science Data Center (ASDC).
- **Outreach**. NAS should seek opportunities for outreach to discover new users and create awareness of the division's capabilities to support big data analysis and analytics on NASA data.

Conclusion

As stated at the beginning of this section, providing the big datasets and/or having the ability to make them available in an environment with the capability to apply a broad range of analysis/analytic algorithms is the key to supporting this user community.

8. Appendices

Appendix I	List of Interviewees
Appendix II	Data Analytics Framework
Appendix III	Interview Questions
Appendix IV	Use Cases

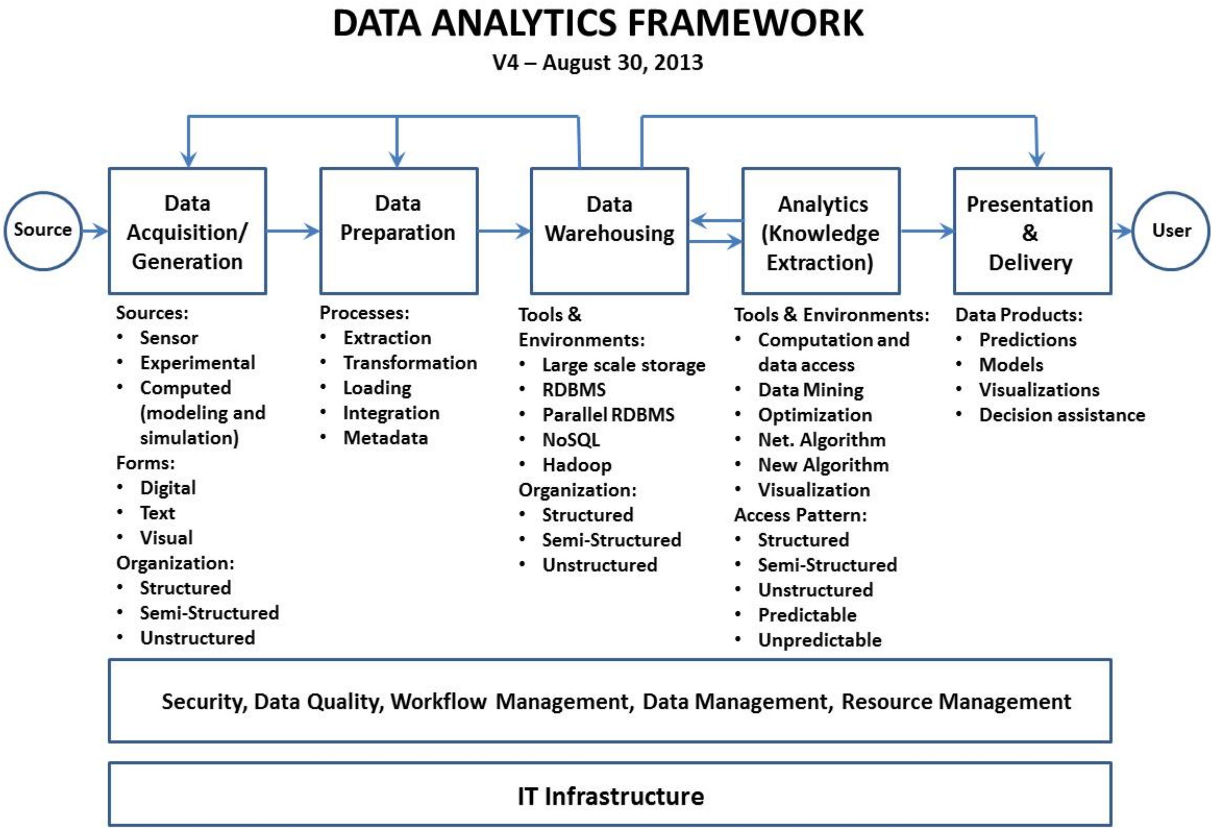
Appendix I - List of Interviewees

The following table identifies the individuals who were interviewed.

Name	Organization	Application
Arlindo Da Silva	Global Modeling and Assimilation Office (GMAO)	Climate modeling
Chris Henze	NAS	Computational physics Signal processing for exoplanet search
Beth Huffer	ASDC	Tools for semantic search and discovery
Dan Kokron	NAS (Former GMAO)	Climate modeling
Vipin Kumar	University of Minnesota	Analytics on Earth science data
Andrew Molthan	Short-term Prediction Research and Transition Center (SPoRT)	Weather modeling
Nikunj Oza	Ames, Intelligent Systems Division	Analytics on flight recorder data Change/anomaly detection using Landsat and MODIS data
Stuart Rogers	NAS	Computational fluid dynamics simulation of flow fields around vehicles
Karen Schleeweis	US Forest Service	Change detection of forested areas
Glenn Tamkin	NCCS	Use of MapReduce on MERRA data
Bridget Thrasher	Climate Analytics Group	Climate downscaling
Petr Votava	NEX	Providing tools and models for deriving properties from satellite data

Appendix II - Data Analysis Framework

The Data Acquisition Framework was developed to provide context for understanding big data analysis and analytics in the data lifecycle from acquisition/generation to ultimate use. The framework helped in formulating the interview questions and in understanding the range of issues involved in providing infrastructure for analysis and analytics.



While the focus of this investigation was on the analysis/analytics per se, the interviews revealed important considerations and involvement for NAS throughout the data lifecycle:

- Simulation/modeling is one of the means of data generation, so NAS is a source as well as a user of big data.
- With missions like Kepler, NAS is involved in data acquisition and data preparation.
- To support analysis/analytics, NAS has to host large datasets locally as well as provide easy means to acquire datasets and move them to NAS.
- Visualization is an important means of making analytic results useful.

Appendix III - Interview Questions

The following questions were used in the interviews.

1. What is your application?

2. What datasets do you use for analysis/analytics?

   2.1. How large are they?

   2.2. Are the data structured or unstructured?

   2.3. Where do they reside? Are they all in one place or multiple places?

   2.4. Do you have to transfer data as part of your application? From where to where?

   2.5. Are any of the datasets you use generated as part of your application (for example, through modeling/simulation)?

      2.5.1. If so, where do you do the computation to generate these datasets?

   2.6. Is there any pre-processing or other preparation needed on the data you analyze before you can do your analysis?

      2.6.1. If so, where is this pre-processing or other preparation done and who does it?

3. Describe the analyses you perform on the data.

   3.1. What is the product/result of your analysis?

      3.1.1. Is this product/result for your own use or does it go to others?

      3.1.2. If to others, how does it get to them?

   3.2. What are the processes/algorithms?

   3.3. What tools/programs do you use to analyze data?

   3.4. Where do you do the analysis (that is, where do you run the tools/programs to analyze data)?

   3.5. Is the analysis you perform generally well defined or is your analysis more exploratory?

   3.6. Are the analyses performed as part of a larger scientific/engineering workflow?

      3.6.1. If so, do you use any tools for workflow management?

4. How would you characterize the size of your data analysis problem? (Things like amount of computation, amount of data, amount of I/O, time it takes, complexity.)

    4.1. Do you have any characterization of the access patterns to data for your application?

5. How is your need for data analysis going to evolve over the next three to five years?

    5.1. What are your unmet needs in this area?

    5.2. How will the size/scale of your needs change? (Characteristics like those in question 4)

    5.3. What additional tools do you need?

6. What could NAS be doing to better support your data analysis needs?

Appendix IV - Use Cases

Six use cases are presented here. For each use case, a user goal is stated, followed by a brief discussion of the use case and possible variations on the use case. This is a preliminary set of use cases based primarily on the interviews. Going forward, it will be useful to elaborate on these and define them in more detail, including quantifying data sizes and computational workload. Also, these use cases tend to reflect the perspective of the current user base. It will be useful to seek out additional use cases from new user communities.

User Goal: Produce a derived dataset by processing NASA data.

The derived dataset could be a large geospatial dataset or it could be a set of parameters or features extracted from the source dataset. The dataset may be available at NAS or the user may have to explicitly move the dataset to NAS from another location. The algorithms may be provided by the user, or they may be made available at NAS. The user may want support in developing the algorithms.

User Goal: Find NASA data relevant to a scientific problem.

The data could reside in any NASA repository at any NASA center. The user needs to be able to articulate the type of data needed and how to tie the scientific problem to the data.

User Goal: Discover new characteristics/features in a NASA dataset.

The dataset may be available at NAS or the user may have to explicitly move the dataset to NAS from another location. A specific characteristic of this use case is that it is exploratory and would typically use different algorithms from analysis use cases—that is, this is at the analytics end of the analysis/analytics continuum.

User Goal: Assess the goodness of a simulation dataset.

In general, this use case assumes the simulation was performed a NAS, but similar cases could utilize data computed at other sites. Additional datasets may be needed to support the assessment. These additional datasets may be available at NAS or the user may have to explicitly move the datasets to NAS from another location. Usually this use case would be at the analysis end of the analysis/analytics continuum.

User Goal: Answer a scientific question through analysis of or analytics on NASA data.

User may need to discover applicable analysis/analytic tools/algorithms. The dataset(s) needed may be available at NAS or the user may have to explicitly move the dataset to NAS from another location. The user may not be an expert on the characteristics of the datasets.

User Goal: Provide the results of analysis/analytics to others.

In many instances, the product of the analytic work is intended for use by a community beyond the user doing the work. This means there is a need to publish/disseminate results, including making derived datasets available to others. These datasets can in some situations be large geospatial datasets.